The Annals of Applied Statistics
2016, Vol. 10, No. 4, 2233–2253
DOI: 10.1214/16-AOAS972

ESTIMATING ODDS RATIOS UNDER A CASE-BACKGROUND DESIGN WITH AN APPLICATION TO A STUDY OF SORAFENIB ACCESSIBILITY

BY JOHN H. SPIVACK AND BIN CHENG

Icahn School of Medicine at Mount Sinai and Columbia University

In certain epidemiologic studies such as those involving stress disorders, sexual harassment, alcohol addiction or epidemiological criminology, exposure data are readily available from cases but not from controls because it is socially inconvenient or even unethical to determine who qualifies as a true control subject. Consequently, it is impractical or even infeasible to use a case-control design to establish the case-exposure association in such situations. To address this issue, we propose a case-background design where in addition to a sample of exposure information from cases, an independent sample of exposure information from the background population is taken, without knowing the case status of the sampled subjects. We develop a semiparametric method to estimate the odds ratio and show that the estimator is strongly consistent and asymptotically normally distributed. Simulation studies indicate that the estimators perform satisfactorily in finite samples and against violations of assumptions. The proposed method is applied to a Sorafenib accessibility study of patients with advanced hepatocellular carcinoma.

1. Introduction.

1.1. *Case-control design and its generalizations*. The case-control design is a primary tool to identify the causes of an effect in epidemiologic research where two independent samples comprising exposure data for cases and controls, called the "case sample" and "control sample" respectively, are obtained retrospectively. The design assumes that if case status is associated with certain exposures, then those exposures should be more prevalent among cases than among controls. An important practical feature of case-control studies is that data sampled retrospectively in a case-control design can be analyzed using a logistic regression model as though they were sampled prospectively in a cohort design [Prentice and Pyke (1979)]. Comprehensive exposition of case-control studies appears in standard references such as those by Breslow and Day (1999) and Keogh and Cox (2014).

The case-control method has been extended to gain efficiency or to decrease cost. One such extension is the nested case-control design proposed by Liddell et al. (1977). A nested case-control design is a single cohort design that prospectively matches emerging cases with concurrently identified and sampled controls

Received November 2015; revised August 2016.
Key words and phrases. Case-only design, criminological epidemiology, disease registry, case-exposure association, imputation, pseudo-likelihood.

to reduce variability and limit the number of controls to be ascertained in order to reduce costs. The design was formally introduced by Prentice and Breslow (1978), and the asymptotic theory associated with this design was developed by Goldstein and Langholz (1992). Another extension is the case-cohort design, or case-base design, where the time to case status is ignored, and hence the outcome is treated as binary. This design, initially proposed by Kupper, McMichael and Spirtas (1975), uses a case sample and a prospectively observed subcohort to increase the yield of cases. Classic references presenting the full development of case-cohort and case-base designs include Miettinen (1976), Prentice (1986) and Nurminen (1989). Asymptotic theory for case-cohort designs was established by Self and Prentice (1988).

1.2. *Sorafenib accessibility study and motivation for new design*. Hepatocellular carcinoma is one of the few malignancies in the United States whose incidence has continued to increase over the past two decades. Despite advances in the field, the 5-year survival remains below 12%, and hepatocellular carcinoma is now the second highest cause of cancer mortality in the world. According to the Barcelona Clinic Liver Cancer staging system, the only official treatment recommendation for patients with Stage C hepatocellular carcinoma is a systemic therapy with oral Sorafenib. The cost of this chemotherapy drug, however, is reported as approximately $5400 per month in the United States [Roberts (2008)]. In addition to the cost, other barriers may also limit the access to this treatment for certain patients, contributing to cancer disparities. A study was undertaken by Heskel et al. (unpublished manuscript) at the Mount Sinai Medical Center, a major urban hospital in New York City, over a 10-year study period to examine access to Sorafenib through physician prescription among eligible hepatocellular carcinoma patients as a function of their sociodemographic information including age, race, socioeconomic status and insurance status. In this study, a total of 352 eligible patients who had access to Sorafenib through prescription were identified by electronic records and their sociodemographic information was collected from the existing electronic data warehouse, which formed a "case sample" of exposure data. However, obtaining a "control sample" of exposure data would require the researchers to identify patients with Stage C hepatocellular carcinoma and confirm that they had not been prescribed Sorafenib before collecting exposures from them. To achieve this would require the investigators to follow patients in the background population through their progression to Stage C and check whether they had received the prescription according to their physicians' notes. The process would require extensive access to personal records, involving review of up to several months of records in paper charts per patient, making it practically infeasible to conduct a traditional case-control study.

The lack of a control sample may happen in many other studies as well. For example, in a study of alcohol addiction, we may be able to acquire exposure information from the cases (i.e., alcohol addicts) from their medical records. However,

to ask a person in a general population whether he or she is an alcohol addict is socially inappropriate, and even if we get an answer to such a query, we cannot expect that answer to be reliable. Similar issues arise in epidemiological criminology where we may have exposure status data for the cases (i.e., convicted criminals), but may be unable to reliably determine who is a control subject among a general population.

In all these situations, a case group with exposure information is easily available, but a control group is not. How then may we estimate the odds ratios of exposure-disease associations in the absence of a control sample? To answer this question, we propose a case-background design which requires a background sample and a prevalence sample in addition to the existing case sample, and prove that the odds ratio can be consistently estimated in this design. To gain an intuitive understanding of this design, consider the simplest case-control data in the form of a 2×2 contingency table. When the control sample is unavailable, this table would be incomplete. However, with the help of a background sample and a prevalence sample, the missing cells can be imputed and the odds ratio can be estimated. A detailed description of the proposed design is given in Section 2, together with an intuitive explanation and full justification of the proposed method to estimate the odds ratio. Simulation results concerning the finite sample performance of the proposed estimator and its sensitivity against violations of assumptions are discussed in Sections 3.1 and 3.2, respectively; the Sorafenib accessibility data are analyzed in Section 3.3; some guidelines for implementing the proposed design are included in Section 3.4; and a brief discussion follows in Section 4. The asymptotic results are proved in the Appendix.

2. Method.

2.1. *Case-background design.* Let Y be a binary random variable indicating the case status with $Y = 1$ denoting a case and $Y = 0$ a control. Let $\boldsymbol{X} = (X_1, \ldots, X_{K-1})^T$ be a $K - 1$-dimensional random vector of exposure variables. In a case-control design, exposure information $\boldsymbol{X}$ is sampled independently from the cases and the controls. The two random samples thus obtained are the case sample and the control sample, respectively, and the case-exposure association may be assessed using a logistic regression model. Unfortunately, as mentioned in Section 1, the situations we consider do not allow for a control sample. There are various reasons that an exposure sample from the controls may be lacking. For instance, in a criminological study, it is easy to obtain exposure information from a case (i.e., a convicted criminal) because we usually have his or her exposure information in the record. On the other hand, it is much more difficult to get such information from a control because it is already challenging to just determine whether a subject in a general population is indeed a control (i.e., has never committed the criminal activity under study), let alone to collect reliable exposure

information from him or her. In the Sorafenib accessibility example, exposure information from the cases (i.e., those who had Stage C hepatocellular carcinoma and had access to Sorafenib) was available in the hospital's electronic registry under a query for Sorafenib prescription followed by a manual review to confirm staging criteria. However, in order to sample exposure data from the controls, one would have to first find patients who were eligible but denied access through prescription and then obtain their exposure information. In our case, although it was not absolutely impossible to obtain an exposure sample from the controls, the cost for doing so was prohibitive, since a query for the lack of a Sorafenib prescription would return large numbers of unrelated records whose manual review to confirm eligibility and staging would exceed available resources. It was thus impractical to use a standard case-control design.

In this paper, we propose a new design, termed a "case-background design," such that under this design we are able to assess the case-exposure association even though a control sample is unavailable. The proposed case-background design requires samples from three independent sources. First, as mentioned above, it requires a case sample of size n_d: the associated exposure information is denoted $\boldsymbol{X}_{d1}, \ldots, \boldsymbol{X}_{dn_d}$, and the case statuses of the subjects in this sample are known as $Y_{di} = 1$, $i = 1, \ldots, n_d$. Second, it requires a random sample, henceforth called the "background sample," of size n_b of the exposure information $\boldsymbol{X}_{b1}, \ldots, \boldsymbol{X}_{bn_b}$, drawn from the background population to which cases belong. The case statuses Y_{bj}, $j = 1, \ldots, n_b$, of the subjects in this sample will not be collected, and hence are unknown. Third, it requires a random sample, henceforth called the "prevalence sample," of size n_p of the case statuses, $Y_{p1}, \ldots, Y_{pn_p}$, from an independent subcohort of the same underlying population, but the exposure information for these subjects, $\boldsymbol{X}_{pl}$, $l = 1, \ldots, n_p$, is not required. We make several remarks concerning the proposed design. First, a background sample is much easier to obtain than a control sample because there is no need to know whether a subject is a case or a control. Second, a prevalence sample, whose size is typically smaller than that of a case sample or a background sample, costs less than a control sample because no exposure information will be collected besides the case status information. Third, the prevalence sample is not needed if an appropriate estimate of the case prevalence in the population under study already exists in the literature.

The intuitive idea of the case-background design is to replace a control sample with a background sample and a prevalence sample. Although the background sample and prevalence sample each carry only partial information—the former has no case status information and the latter has no exposure information—in combination they can play a similar role to a control sample in estimating the odds ratio. The design is therefore useful when three criteria are met. First, a case sample already exists or can be obtained easily; second, a control sample is impossible, unethical or costly; and third, a background sample and a prevalence sample can be obtained at reasonable cost.

The Sorafenib accessibility study met these criteria. First, the case sample was of eligible patients who had access to Sorafenib through prescription. They were identified by a query of electronic records followed by a manual review of the clinical stage, after which sociodemographic information was collected from the existing electronic data warehouse. Second, it was burdensome due to the form of the database and details of the staging system to acquire a sufficiently large sample of patients who met the Stage C classification and had not received a Sorafenib prescription. Third, a background sample and a prevalence sample could be obtained within the limits of resource constraints. Specifically, the electronic records allowed the identification of patients who were at an earlier, pre-metastatic stage. Although these patients themselves were not eligible for Sorafenib, it was considered clinically valid to assume that the sociodemographic information of this pre-metastatic population did not change drastically over the short time period during which they would progress to become eligible for Sorafenib. These patients were considered as a background sample. Clearly, it is impossible to know the case statuses of subjects in the background sample without lengthy follow-up, which favors a strategy where case status information is not collected from the background sample. A prevalence sample of eligible patients was identified by a manual chart review and their prescription statuses determined. However, the logistical burden to identify these patients and collect their sociodemographic information was heavy, leading to the smaller size of the prevalence sample and supporting the choice not to fully collect exposure information on the prevalence sample. It is clear that for this application the proposed case-background design had a clear advantage in logistical burden and cost effectiveness.

2.2. *Model.* We assume that the joint distribution of case status and exposure information $(Y, \boldsymbol{X}^T)$ is specified hierarchically such that the marginal distribution of $\boldsymbol{X}$ has a probability density function $f(\boldsymbol{x})$ with respect to a certain σ-finite measure μ on the $K-1$-dimensional Euclidean space, and, given $\boldsymbol{X}$, the conditional distribution of Y satisfies a logistic regression model

$$\log\left\{\frac{\operatorname{pr}(Y=1 \mid \boldsymbol{X})}{1-\operatorname{pr}(Y=1 \mid \boldsymbol{X})}\right\} = \beta_0 + \sum_{j=1}^{K-1} x_j \beta_j = \tilde{\boldsymbol{X}}^T \boldsymbol{\beta}, \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{K-1})^T$ and $\tilde{\boldsymbol{X}} = \{1, \boldsymbol{X}^T\}^T$ is the design vector with constant 1 added to $\boldsymbol{X}$ for the intercept β_0.

We also assume that the case sample, the background sample and the prevalence sample required in the case-background design are from a common distribution $(Y, \boldsymbol{X}^T)$ described above; that is, we assume that the case sample is a random sample from the conditional distribution of $\boldsymbol{X}$ given $Y = 1$ whose probability density function is denoted as $f(\boldsymbol{x} \mid Y = 1)$; the background sample is a random sample from the marginal distribution of $\boldsymbol{X}$ whose probability density function is $f(\boldsymbol{x})$; and the prevalence sample is a random sample from the marginal distribution of

Y, a Bernoulli distribution with parameter $\pi_d = \text{pr}(Y = 1)$. The objective of this paper is to estimate $\boldsymbol{\beta}$, particularly $\beta_1, \ldots, \beta_{K-1}$, based on these three samples.

2.3. *Estimating method.* To motivate the proposed estimating method, consider a simple logistic model where there is one single dichotomized exposure variable X with $X = 1$ denoting the exposed and $X = 0$ the unexposed. Let n_d^+ and n_d^- be the numbers of the exposed and the unexposed subjects in the case sample. Then data from this case sample yields an incomplete 2×2 contingency table.

If the missing cells can be appropriately imputed, then we can estimate the odds ratio based on the imputed contingency table by treating it as one from a genuine case-control design. We now describe how to impute the missing cells in Table 1. Let $\hat{\pi}_d$ denote the estimated case prevalence based on the prevalence sample. Then the overall total in the above table can be imputed as $n_d/\hat{\pi}_d$. From the background sample we estimate the exposure rate as n_b^+/n_b, where n_b^+ denotes the number of exposed subjects out of a total of n_b subjects in the background sample. Similarly, n_b^- denotes the number of unexposed subjects out of the n_b subjects in the background sample. Therefore, out of an imputed overall total of $n_d/\hat{\pi}_d$ subjects, $(n_d/\hat{\pi}_d)(n_b^+/n_b)$ subjects should be exposed. As the result, we obtain an imputed contingency table.

Denote

$$n_c^+ = \frac{n_d}{\hat{\pi}_d}\frac{n_b^+}{n_b} - n_d^+, \qquad n_c^- = \frac{n_d}{\hat{\pi}_d}\frac{n_b^-}{n_b} - n_d^-, \qquad n_c = n_c^+ + n_c^- = \frac{n_d}{\hat{\pi}_d} - n_d.$$

Then n_c^+ and n_c^- can be viewed as the numbers of exposed and unexposed subjects in a "pseudo" control group of n_c subjects, respectively. A natural estimate of the odds ratio between case and exposure from Table 2 is

$$\widehat{\text{OR}} = \frac{n_d^+ n_c^-}{n_d^- n_c^+}. \tag{2.2}$$

It can be directly shown that this heuristic estimate is strongly consistent for the true odds ratio and has an asymptotically normal distribution.

TABLE 1
Incomplete contingency table based on the case sample only

	$Y = 1$	$Y = 0$	Total
$X = 1$	n_d^+	?	?
$X = 0$	n_d^-	?	?
Total	n_d	?	?

TABLE 2
Imputed contingency table based on the case sample, the background sample and the prevalence sample

	$Y=1$	$Y=0$	Total
$X=1$	n_d^+	$(n_d/\hat{\pi}_d)(n_b^+/n_b)-n_d^+$	$(n_d/\hat{\pi}_d)(n_b^+/n_b)$
$X=0$	n_d^-	$(n_d/\hat{\pi}_d)(n_b^-/n_b)-n_d^-$	$(n_d/\hat{\pi}_d)(n_b^-/n_b)$
Total	n_d	$n_d/\hat{\pi}_d-n_d$	$n_d/\hat{\pi}_d$

To estimate the odds ratio consistently, it is important to use consistent estimators for both $\mathrm{pr}(X=1 \mid Y=1)$ and $\mathrm{pr}(X=1 \mid Y=0)$. We estimate the former by n_d^+/n_d based on the case sample and the latter by n_c^+/n_c based on a pseudo-control sample, and both estimators are consistent. Using other estimators, such as $(n_d^+ + n_c^+)/(n_d+n_c)$ or n_b^+/n_b, which estimate $\mathrm{pr}(X=1)$ instead of $\mathrm{pr}(X=1 \mid Y=0)$, will lead to inconsistent estimation of the odds ratio. Practically, it is only appropriate to use these inconsistent estimators when cases are very rare. In our example, however, the prevalence of prescription of Sorafenib in the population was close to 50%. Therefore, using the above estimators would lead to inconsistent estimates.

To generalize the imputed contingency table idea to the situation where there are multiple exposure variables of either categorical or continuous type, we note that, according to known results [Prentice and Pyke (1979)], the odds ratio estimator can be equivalently obtained by exponentiating the maximum likelihood estimator of β_1 in the following simple logistic regression model with a single binary covariate X indicating the exposure status

$$\log\left\{\frac{\mathrm{pr}(Y=1 \mid X)}{1-\mathrm{pr}(Y=1 \mid X)}\right\} = \beta_0 + \beta_1 X.$$

Therefore, we should be able to obtain the estimated odds ratio given in (2.2) by exponentiating the maximum likelihood estimator of β_1 treating the imputed contingency table as one from a genuine case-control study. After some algebraic manipulations, the pseudo-log-likelihood function associated with Table 2 can be expressed as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n_d}(\beta_0+\beta_1 X_{di}) - \frac{n_d}{n_b\hat{\pi}_d}\sum_{j=1}^{n_b}\log\bigl(1+e^{\beta_0+\beta_1 X_{bj}}\bigr).$$

The above function is called a pseudo-log-likelihood because it is based on pseudo-data in Table 2, not on the actual data. On the other hand, the above heuristic argument does suggest that, for a multiple logistic regression model (2.1) where there are $K-1$ exposure variables, the coefficient vector $\boldsymbol{\beta}$ can be estimated by

the maximizer of the pseudo-log-likelihood function

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n_d} \tilde{\boldsymbol{X}}_{di}^T \boldsymbol{\beta} - \frac{n_d}{n_b \hat{\pi}_d} \sum_{j=1}^{n_b} \log\bigl(1 + e^{\tilde{\boldsymbol{X}}_{bj}^T \boldsymbol{\beta}}\bigr)$$

or, equivalently, the solution to the pseudo-score functions

$$\boldsymbol{s}(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n_d} \tilde{\boldsymbol{X}}_{di} - \frac{n_d}{n_b \hat{\pi}_d} \sum_{j=1}^{n_b} \frac{\tilde{\boldsymbol{X}}_{bj} e^{\tilde{\boldsymbol{X}}_{bj}^T \beta}}{1 + e^{\tilde{\boldsymbol{X}}_{bj}^T \beta}} = \boldsymbol{0}. \tag{2.3}$$

The proposed method is essentially an estimating equation method. However, unlike the usual estimating equations which are derived based on the actual data, our proposed estimating equation is from a pseudo-score of the imputed data. Consequently, the asymptotic properties of the proposed estimator must be reestablished. In the Appendix we prove that the proposed method yields a consistent estimator of $\boldsymbol{\beta}$. Specifically, we have the following result.

THEOREM 2.1. *Let $n = n_d + n_b + n_p$. Assume that $\lim_{n\to\infty} n_i/n = \rho_i$ exists and satisfies $0 < \rho_i < 1$, $i = d, b, p$. With probability approaching 1 as $n \to \infty$, the solution to (2.3), denoted as $\hat{\boldsymbol{\beta}}$, exists and is unique. Let $\boldsymbol{\beta}^*$ be the true unknown value for model (2.1). Assume that $\operatorname{var}(\boldsymbol{X})$ is positive definite. Then, as $n \to \infty$, $\hat{\boldsymbol{\beta}} \to_{a.s.} \boldsymbol{\beta}^*$, and*

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \to_d N_K\{\boldsymbol{0}, \boldsymbol{M}(\boldsymbol{\beta}^*)^{-1} \boldsymbol{V} \boldsymbol{M}(\boldsymbol{\beta}^*)^{-1}\},$$

where

$$\boldsymbol{V} = \rho_d^{-1} \pi_d^2 \operatorname{var}(\tilde{\boldsymbol{X}} \mid Y = 1) + \rho_b^{-1} \operatorname{var}\biggl(\frac{\tilde{\boldsymbol{X}} e^{\tilde{\boldsymbol{X}}^T \boldsymbol{\beta}^*}}{1 + e^{\tilde{\boldsymbol{X}}^T \boldsymbol{\beta}^*}}\biggr)$$
$$+ \rho_p^{-1} \pi_d (1 - \pi_d) E(\tilde{\boldsymbol{X}} \mid Y = 1)\bigl\{E(\tilde{\boldsymbol{X}} \mid Y = 1)\bigr\}^T,$$

and $\boldsymbol{M}(\boldsymbol{\beta}^*)$ is a $K \times K$ positive definite matrix whose (k, l)th entry is

$$\int x_{k-1} x_{l-1} e^{\tilde{\boldsymbol{x}}^T \boldsymbol{\beta}^*} \bigl(1 + e^{\tilde{\boldsymbol{x}}^T \boldsymbol{\beta}^*}\bigr)^{-2} f(\boldsymbol{x})\, d\mu(\boldsymbol{x}) \qquad (k, l = 1, \ldots, K),$$

where $x_0 \equiv 1$ and x_{k-1} is the $(k-1)$th element of exposure vector $\boldsymbol{x}$.

The matrix $\boldsymbol{V}$ can be consistently estimated by plugging in the following estimators:

$$\hat{\pi}_d = \frac{1}{n_p} \sum_{l=1}^{n_p} I(Y_{pl} = 1), \qquad \hat{E}(\tilde{\boldsymbol{X}} \mid Y = 1) = \frac{1}{n_d} \sum_{i=1}^{n_d} \tilde{\boldsymbol{X}}_{di},$$

where $I(\cdot)$ is the indicator function, and

$$\widehat{\operatorname{var}}(\tilde{\boldsymbol{X}} \mid Y = 1) = \frac{1}{n_d - 1} \sum_{i=1}^{n_d} (\tilde{\boldsymbol{X}}_{di} - \overline{\tilde{\boldsymbol{X}}}_d)(\tilde{\boldsymbol{X}}_{di} - \overline{\tilde{\boldsymbol{X}}}_d)^T,$$

where $\overline{\tilde{\boldsymbol{X}}}_d$ denotes the sample mean of $\tilde{\boldsymbol{X}}_{di}$, $i = 1, \ldots, n_d$, and

$$\widehat{\text{var}}\left(\frac{\tilde{\boldsymbol{X}} e^{\tilde{\boldsymbol{X}}^T \beta^*}}{1 + e^{\tilde{\boldsymbol{X}}^T \beta^*}}\right) = \frac{1}{n_b - 1} \sum_{j=1}^{n_b} (\boldsymbol{W}_{bj} - \overline{\boldsymbol{W}}_b)(\boldsymbol{W}_{bj} - \overline{\boldsymbol{W}}_b)^T,$$

where

$$\boldsymbol{W}_{bj} = \tilde{\boldsymbol{X}}_{bj} e^{\tilde{\boldsymbol{X}}_{bj}^T \hat{\boldsymbol{\beta}}} \big(1 + e^{\tilde{\boldsymbol{X}}_{bj}^T \hat{\boldsymbol{\beta}}}\big)^{-1} \qquad (j = 1, \ldots, n_b),$$

and $\overline{\boldsymbol{W}}_b$ is the sample mean of $\boldsymbol{W}_{bj}$, $j = 1, \ldots, n_b$. Finally, $M(\boldsymbol{\beta}^*)$ is estimated consistently by

$$\hat{\boldsymbol{M}}(\boldsymbol{\beta}^*) = \frac{1}{n_b} \sum_{j=1}^{n_b} \tilde{\boldsymbol{X}}_{bj} \tilde{\boldsymbol{X}}_{bj}^T e^{\tilde{\boldsymbol{X}}_{bj}^T \hat{\boldsymbol{\beta}}} \big(1 + e^{\tilde{\boldsymbol{X}}_{bj}^T \hat{\boldsymbol{\beta}}}\big)^{-2}.$$

We remark that, although the intended applications of the proposed design are to studies for which a control sample is not available, including a control sample, in case it can be obtained by some means, into the new design will generally improve the efficiency. To see this, suppose in addition to n_d case sample subjects, n_b background sample subjects and n_p prevalence sample subjects, we also have n_c control sample subjects, whose exposures are denoted as $\boldsymbol{X}_{c1}, \ldots, \boldsymbol{X}_{c,n_c}$. Then parameter $\boldsymbol{\gamma} = E(\tilde{\boldsymbol{X}} \mid Y = 1)$, which appears in the estimating equation in the Appendix and is estimated by $\frac{1}{n_d} \sum_{i=1}^{n_d} \tilde{\boldsymbol{X}}_{di}$ in the case-background design, can now be better estimated using both $\frac{1}{n_d} \sum_{i=1}^{n_d} \tilde{\boldsymbol{X}}_{di}$ and

$$\frac{\frac{1}{n_b} \sum_{j=1}^{n_b} \tilde{\boldsymbol{X}}_{bj} - (\frac{1}{n_c} \sum_{k=1}^{n_c} \tilde{\boldsymbol{X}}_{ck}) \{\frac{1}{n_p} \sum_{l=1}^{n_p} I(Y_{pl} = 0)\}}{\frac{1}{n_p} \sum_{l=1}^{n_p} I(Y_{pl} = 1)},$$

which leads to improved efficiency. The magnitude of the efficiency improvement, however, is hard to determine, as it depends on other unknown parameters as well.

3. Numerical studies.

3.1. *Finite sample performance.* In this section, we conduct simulation studies to evaluate the performance of the proposed estimator. The simulation setting is similar to but simpler than the one in Section 3.3 for the Sorafenib accessibility study. The binary outcome Y is the accessibility of Sorafenib among eligible patients, which has the role of "case status" in our notation, with $Y = 1$ indicating that a patient has access to Sorafenib, and hence is considered a "case," and $Y = 0$ indicating that a patient, although eligible, does not have access to Sorafenib, and hence is a "control." Two dichotomous exposure variables, socioeconomic status (SES, high versus low) and race (nonwhite versus white), taken from the collection of exposure variables used in the Sorafenib study, are simulated from

TABLE 3
Joint distribution of socioeconomic status and race in background and case populations. SES: socioeconomic status

	Background population		**Case population**	
	SES = high	**SES = low**	**SES = high**	**SES = low**
Race = nonwhite	0.209	0.495	0.260	0.402
Race = white	0.205	0.090	0.262	0.076

a multinomial distribution whose cell probabilities, displayed in the second and third columns of Table 3, are based on their joint distribution in the Sorafenib accessibility study. Then, conditioning on SES and race, case status Y is generated from a Bernoulli distribution whose probability is given via the following logistic regression model:

$$\log\left\{\frac{\text{pr}(Y_i = 1 \mid \text{SES}_i, \text{race}_i)}{1 - \text{pr}(Y_i = 1 \mid \text{SES}_i, \text{race}_i)}\right\} = \beta_0 + \beta_1 \text{SES}_i + \beta_2 \text{race}_i, \tag{3.1}$$

where the true parameter values $\beta_0 = -0.750$, $\beta_1 = 0.700$ and $\beta_2 = -0.050$, are set equal to those estimated from the Sorafenib example. From Table 3 and the logistic regression model (3.1), the case prevalence in the simulation is determined as 0.382 by the total probability theorem, and the joint distribution of SES and race in the case population is determined as in the forth and fifth columns of Table 3 by Bayes' theorem. Note that we choose the true β_i, $i = 0, 1, 2$, to be the estimates derived from the Sorafenib accessibility example, and that since the example contains two more covariates, the prevalence rate in the simulation differs from that of the example.

Three choices of total sample size, $n = 500$, 1000 and 1500, and various choices of partition of the total sample size into the three subsamples, (ρ_d, ρ_b, ρ_p), are considered. We choose ρ_d and ρ_p to be relatively small compared with ρ_b, to reflect the fact that in actual studies where the case-background design is applicable, it is usually much easier and cheaper to sample from the background than from the case group or the prevalence group. The expectation of $\hat{\beta}_i$, the standard error $\text{se}(\hat{\beta}_i)$ and the coverage probability of the 95% asymptotic confidence interval for β_i, $i = 0, 1, 2$, are estimated based on 5000 simulations. The results are summarized in Table 4.

The simulation results indicate that the coverage probabilities of the asymptotic confidence intervals are slightly greater than the nominal level of 95%, but approach 95% as the total sample size increases. When the total sample size n is fixed, the simulation results suggest that, although it is cheaper to obtain the background sample, it is not true that the higher the proportion of the background sample ρ_b, the better the performance; that is, the proportions of the case and the

TABLE 4

Summary of simulation results. True parameters $(\beta_0, \beta_1, \beta_2) = (-0.750, 0.700, -0.050)$. *SE*: *empirical standard error*; *CP*: *empirical coverage probability*

		β_0			β_1			β_2		
n	(ρ_d, ρ_b, ρ_p)	**Bias**	**SE**	**CP**	**Bias**	**SE**	**CP**	**Bias**	**SE**	**CP**
500	(0.2, 0.6, 0.2)	−0.016	0.503	0.969	0.036	0.462	0.966	0.012	0.506	0.977
500	(0.3, 0.6, 0.1)	0.006	0.480	0.968	0.030	0.413	0.967	−0.018	0.431	0.977
500	(0.2, 0.7, 0.1)	−0.016	0.539	0.970	0.029	0.460	0.971	−0.003	0.506	0.980
500	(0.1, 0.6, 0.3)	−0.017	0.648	0.984	0.038	0.601	0.976	0.001	0.675	0.978
500	(0.1, 0.7, 0.2)	−0.015	0.652	0.983	0.041	0.600	0.974	−0.004	0.670	0.977
1000	(0.2, 0.6, 0.2)	−0.008	0.339	0.962	0.009	0.310	0.961	0.004	0.341	0.962
1000	(0.3, 0.6, 0.1)	−0.007	0.341	0.951	0.014	0.275	0.961	0.000	0.299	0.959
1000	(0.2, 0.7, 0.1)	−0.022	0.367	0.951	0.014	0.307	0.958	0.006	0.330	0.963
1000	(0.1, 0.6, 0.3)	−0.007	0.423	0.970	0.022	0.404	0.962	−0.003	0.443	0.968
1000	(0.1, 0.7, 0.2)	−0.017	0.433	0.964	0.024	0.407	0.960	0.007	0.448	0.961
1500	(0.2, 0.6, 0.2)	−0.005	0.277	0.952	0.012	0.250	0.959	0.001	0.277	0.955
1500	(0.3, 0.6, 0.1)	−0.010	0.271	0.950	0.011	0.222	0.958	0.005	0.237	0.959
1500	(0.2, 0.7, 0.1)	−0.002	0.284	0.960	0.011	0.244	0.959	−0.004	0.263	0.961
1500	(0.1, 0.6, 0.3)	−0.011	0.343	0.963	0.015	0.331	0.956	0.004	0.357	0.963
1500	(0.1, 0.7, 0.2)	−0.005	0.352	0.957	0.007	0.331	0.954	−0.002	0.357	0.957

prevalence samples, ρ_d and ρ_p, cannot be chosen too small. When the total sample size n and the background sample proportion ρ_b are fixed, having a bigger case sample, that is, $\rho_d > \rho_p$, generally results in better performance than having a bigger prevalence sample, that is, $\rho_p > \rho_d$. Empirically, it seems that a $3:6:1$ ratio in the case, the background and the prevalence sample sizes or, equivalently, the partition $(\rho_d, \rho_b, \rho_p) = (0.3, 0.6, 0.1)$, yields favorable performance. When the total sample size n is smaller than 500, the performance begins to deteriorate and histograms of the distributions of parameter estimators show departures from normality. This suggests that, for smaller sample sizes, the asymptotic results may not be preferred and alternatives such as bootstrap confidence intervals may be a superior. However, we do not explore such proposals in this paper.

3.2. *Sensitivity analysis*. The key assumption needed for the case-background design is that the case sample, the background sample and the prevalence sample derive from the same underlying population. When this assumption is violated, bias may occur, and if the violation is severe, the proposed method may not even be valid. In this section, we investigate the impact of violations of this assumption via a set of simulation studies. Specifically, 18 scenarios are considered (see Tables 5 and 6). The first 8 scenarios correspond to 8 perturbations of the exposure distribution in the background population given in Table 3. They represent the situations where the background sample is taken from a population slightly different from the underlying population from which cases arise. The perturbations are formed by adding or subtracting a fixed percentage of the true value from a cell probability in Table 3 and compensating the remaining three cell probabilities evenly in order to maintain the unity total probability constraint. For example, in the first set of simulations using 5% relative perturbations, Scenario 1, $\pi_{11}^b \uparrow$, corresponds to adding $0.010 = (0.050)(0.209)$ to the top left cell probability 0.209 and subtracting $0.003 = (0.50)(0.209)/3$ from the remaining cell probabilities of the table. Similarly, Scenario 2, $\pi_{11}^b \downarrow$, corresponds to subtracting 0.010 to the top left cell probability 0.209 and adding 0.003 to all the rest. Scenarios 9 to 16 represent 8 perturbations of the exposure distribution in the case population in Table 3, and the perturbation procedure is the same as for the first 8 scenarios. The last two scenarios correspond to increasing and decreasing the case prevalence rate, respectively, and represent situations where the prevalence sample is taken from a population that differs slightly from the underlying population where cases arise. The first set of simulations shows effects of 5% relative perturbations; the second set shows effects of 10% relative perturbations considered as violations of the design assumptions.

The simulations are done under the parameters $n = 1500$, $(\rho_d, \rho_b, \rho_p) = (0.2, 0.7, 0.1)$, approximating those of the Sorafenib example. All estimations are based on 5000 simulation runs, and results are summarized in Tables 5 and 6 for 5% and 10% perturbations, respectively. Violations of assumptions lead to increased standard errors across all the 18 scenarios and for both levels of perturbation, but the coverage probabilities appear unaffected by 5% relative perturbations

TABLE 5

Summary of sensitivity analysis results: *version* 1. ↑ (*resp.* ↓): *add* (*resp. subtract*) 5% *of the probability of the specified cell and redistribute evenly to the remaining cells*; π_{ij}^{b}: *exposure prevalence in background population*; π_{ij}^{d}: *exposure prevalence in case population*, $i = 1$ (*nonwhite*), 2 (*white*), $j = 1$ (*high SES*), 2 (*low SES*); π_{d}^{p}: *case prevalence in prevalence population*; $n = 1500$ *and* $(\rho_d, \rho_b, \rho_p) = (0.2, 0.7, 0.1)$. *True parameters* $(\beta_0, \beta_1, \beta_2) = (-0.750, 0.700, -0.050)$; *SE*: *empirical standard error*; *CP*: *empirical coverage probability*

		β_0			β_1			β_2		
Scenario	**Perturbation**	**Bias**	**SE**	**CP**	**Bias**	**SE**	**CP**	**Bias**	**SE**	**CP**
1	π_{11}^{b} ↑	0.090	0.354	0.947	−0.068	0.324	0.945	−0.091	0.360	0.953
2	π_{11}^{b} ↓	−0.120	0.356	0.948	0.108	0.338	0.957	0.108	0.365	0.953
3	π_{21}^{b} ↑	−0.043	0.351	0.963	−0.012	0.329	0.958	0.066	0.357	0.958
4	π_{21}^{b} ↓	0.023	0.344	0.959	0.042	0.319	0.960	−0.056	0.356	0.954
5	π_{12}^{b} ↑	0.029	0.375	0.961	0.100	0.345	0.948	−0.090	0.376	0.955
6	π_{12}^{b} ↓	−0.039	0.333	0.958	−0.080	0.322	0.940	0.087	0.345	0.948
7	π_{22}^{b} ↑	−0.037	0.340	0.962	0.041	0.328	0.956	0.029	0.352	0.958
8	π_{22}^{b} ↓	0.017	0.349	0.962	−0.012	0.326	0.957	−0.020	0.362	0.956
9	π_{11}^{d} ↑	−0.126	0.352	0.950	0.106	0.332	0.955	0.109	0.364	0.949
10	π_{11}^{d} ↓	0.105	0.348	0.949	−0.083	0.320	0.947	−0.100	0.356	0.949
11	π_{21}^{d} ↑	0.006	0.351	0.957	0.055	0.327	0.959	−0.046	0.363	0.952
12	π_{21}^{d} ↓	−0.033	0.349	0.959	−0.026	0.326	0.953	0.059	0.356	0.957
13	π_{12}^{d} ↑	−0.036	0.341	0.959	−0.048	0.327	0.947	0.077	0.351	0.953
14	π_{12}^{d} ↓	0.015	0.360	0.958	0.085	0.329	0.960	−0.077	0.368	0.950
15	π_{22}^{d} ↑	0.028	0.352	0.960	−0.019	0.327	0.953	−0.032	0.361	0.956
16	π_{22}^{d} ↓	−0.042	0.349	0.957	0.044	0.326	0.960	0.033	0.361	0.958
17	π_{d}^{p} ↑	0.070	0.364	0.952	0.038	0.343	0.960	−0.004	0.379	0.955
18	π_{d}^{p} ↓	−0.078	0.338	0.955	−0.015	0.315	0.954	0.003	0.346	0.956

TABLE 6

Summary of sensitivity analysis results: *version* 2. ↑ (*resp.* ↓): *add* (*resp. subtract*) 10% *of the probability of the specified cell and redistribute evenly to the remaining cells*; π_{ij}^b: *exposure prevalence in background population*; π_{ij}^d: *exposure prevalence in case population*, $i = 1$ (*nonwhite*), 2 (*white*), $j = 1$ (*high SES*), 2 (*low SES*); π_d^p: *case prevalence in prevalence population*; $n = 1500$ *and* $(\rho_d, \rho_b, \rho_p) = (0.2, 0.7, 0.1)$. *True parameters* $(\beta_0, \beta_1, \beta_2) = (-0.750, 0.700, -0.050)$; *SE*: *empirical standard error*; *CP*: *empirical coverage probability*

		β_0			β_1			β_2		
Scenario	**Perturbation**	**Bias**	**SE**	**CP**	**Bias**	**SE**	**CP**	**Bias**	**SE**	**CP**
1	$\pi_{11}^b \uparrow$	0.193	0.356	0. 926	−0.147	0.318	0.918	−0.191	0.364	0.928
2	$\pi_{11}^b \downarrow$	−0.222	0.355	0. 924	0.197	0.349	0.939	0.211	0.364	0.934
3	$\pi_{21}^b \uparrow$	−0.080	0.365	0. 956	−0.038	0.342	0.945	0.127	0.363	0.946
4	$\pi_{21}^b \downarrow$	0.066	0.339	0. 955	0.070	0.318	0.961	−0.125	0.357	0.947
5	$\pi_{12}^b \uparrow$	0.071	0.390	0. 964	0.193	0.363	0.934	−0.176	0.396	0.944
6	$\pi_{12}^b \downarrow$	−0.047	0.312	0. 959	−0.180	0.305	0.907	0.158	0.327	0.927
7	$\pi_{22}^b \uparrow$	−0.078	0.339	0. 954	0.076	0.333	0.949	0.033	0.349	0.959
8	$\pi_{22}^b \downarrow$	0.060	0.355	0. 964	−0.059	0.329	0.948	−0.056	0.361	0.959
9	$\pi_{11}^d \uparrow$	−0.249	0.362	0. 908	0.203	0.339	0.932	0.226	0.370	0.920
10	$\pi_{11}^d \downarrow$	0.218	0.347	0. 913	−0.179	0.320	0.908	−0.201	0.360	0.918
11	$\pi_{21}^d \uparrow$	0.028	0.345	0. 955	0.096	0.325	0.953	−0.107	0.357	0.947
12	$\pi_{21}^d \downarrow$	−0.051	0.360	0.960	−0.069	0.331	0.949	0.115	0.367	0.942
13	$\pi_{12}^d \uparrow$	−0.063	0.334	0. 956	−0.115	0.319	0.933	0.164	0.344	0.937
14	$\pi_{12}^d \downarrow$	0.049	0.367	0. 960	0.139	0.333	0.948	−0.163	0.378	0.933
15	$\pi_{22}^d \uparrow$	0.058	0.354	0. 953	−0.039	0.333	0.950	−0.054	0.362	0.954
16	$\pi_{22}^d \downarrow$	−0.061	0.345	0.957	0.059	0.324	0.962	0.049	0.355	0.961
17	$\pi_d^p \uparrow$	0.130	0.378	0. 946	0.063	0.358	0.959	0.008	0.394	0.959
18	$\pi_d^p \downarrow$	−0.157	0.331	0. 930	−0.040	0.301	0.953	0.002	0.337	0.955

and show a mild impact from 10% relative perturbations. Perturbations do incur positive or negative bias in the point estimators, depending on the direction of the perturbation, but the bias is generally modest in units of the standard error. In summary, the proposed method shows satisfactory robustness when the assumption that all samples originate from the same underlying population is subject to mild violations. On the other hand, as in any epidemiological study, caution must be exercised when sampling the data to ensure the main assumption of the design is approximately met. Some practical guidelines for responsible use of this design are provided in Section 3.4.

3.3. *Example*: *Sorafenib accessibility study.* The study collected a case sample of 352 patients, a background sample of 959 patients and a prevalence sample of 137 patients which yielded a case prevalence rate of 45.3%. Interest was focused on assessing how the four predictors: age ($\geq$65 versus $<$65), race (nonwhite versus white), socioeconomic status (high versus low by the median income of a patient's zip code of residence) and insurance (government alone versus commercial or government plus supplement) affected access to Sorafenib. The odds ratios, 95% asymptotic confidence intervals and p-values for the four predictors were calculated by the proposed method using the three samples. The results, which are summarized in Table 7, reveal that patients of high socioeconomic status had significantly greater access to Sorafenib; older age was somewhat associated with improved access, but the effect did not reach significance. On the other hand, neither race nor insurance type significantly influenced Sorafenib accessibility.

To summarize, although it was not feasible to obtain a control sample for the Sorafenib study, with the help of a background sample and a prevalence sample, it was still possible to assess the case-exposure associations. Importantly, the proposed method allowed us to detect a disparity in access to Sorafenib between patients with different socioeconomic statuses at a major urban medical center that experienced continuity of its internal operations, populations served and external outreach during the 10-year study period. There are several possible explanations for the observed disparity. One is that patients of high SES may be less deterred

TABLE 7
Estimated odds ratios for exposure variables with 95% confidence intervals and p-values. OR: odds ratio; Asymp CI: asymptotic confidence interval; SES: socioeconomic status

Exposure variable	$\widehat{\mathbf{OR}}$	**95% Asym CI**	***p*-value**
Age ($\geq$65 versus $<$65)	1.535	0.921–2.560	0.100
Race (nonwhite versus white)	0.947	0.520–1.725	0.858
SES (high versus low)	2.073	1.191–3.608	0.010
Insurance (government alone versus commercial or government plus supplement)	1.202	0. 745–1.940	0.451

by costs such as copayments or more able to accept out-of-pocket expenses when facing obstacles in financing their care through insurance. Another possible reason is that patients of high SES are typically better educated, more knowledgeable about their treatment options and more likely to demand access to care. Other explanations could involve the strength of patients' support networks and levels of family involvement. Finally, it is possible that high SES patients had access to doctors who were more knowledgeable and more willing to share their treatment knowledge and provide treatment options. The detected disparity should be further researched to confirm its generalizability beyond the single academic medical center where the study was conducted. Nonetheless, there are potentially important implications for policy making; more research is needed to identify barriers that lead to disparities in care and more resources should be directed toward socioeconomically disadvantaged patients to improve their access to state-of-the-art therapies. Our finding also signifies the importance of the new proposed method because in the absence of a control sample such a disparity would most likely go unnoticed, in spite of ongoing efforts to improve the quality and comprehensiveness of institutional electronic medical records.

3.4. *Practical guidelines.* We expect the proposed design to be embraced by applied researchers who have ready access to a case sample, for example, through electronic medical records, but have difficulty obtaining a control sample due to either ethical or budgetary restrictions, such that they are unable to conduct a usual case-control design. In such studies, a case-background design has a clear advantage over the traditional case-control design in terms of efficiency and cost, and thus should be considered. On the other hand, implementing a case-background design requires caution to ensure that all samples reflect the same underlying conceptual population since inappropriate sampling would lead to bias and misleading findings. In this section we propose some practical guidelines to ensure proper utilization of the proposed method.

First, at the planning stage, an underlying population should be identified through specific inclusion and exclusion criteria, and the definition of a case (and hence of a control) should be clearly stated. This requirement seems more important for the case-background design because in reality investigators might already have a case sample in hand when they think of using the case-background design, and they would have to think carefully about what they intend as their underlying population before taking a background sample. For example, in the Sorafenib accessibility study, only a case sample through a disease registry was available at the beginning of the study, and we decided to sample the background population only when we realized that sampling the controls was nearly impossible. In such a situation, bias will easily occur if one is not clear about what underlying population one is referring to and what is meant by a subject being a case.

Second, if time trends present a possible source of bias in a given study design, one should choose to measure exposures in ways that are temporally stable to make sure the exposure distribution in the background sample is as close as possible to the one in the underlying population to which the cases belong. This way, any temporal change in the underlying population will have lesser impact on the exposure distribution. In the Sorafenib study, the socioeconomic status was measured by the median income of a patient's zip code of residence, which is clearly more stable and robust than the patient's annual income, as the latter may have great fluctuations. This recommendation is more relevant if there is a temporal gap between the background population and the underlying population for cases. For example, in the Sorafenib study, the background population consisted of patients who were at a pre-metastatic stage and needed a short time period to progress to stage C.

Third, in taking the prevalence sample, choose a sampling scheme as unbiased as possible and a case status determination as objective as possible to ensure the prevalence sample correctly reflects the case prevalence in the underlying population. As we have emphasized, identifying who are the cases and who are the controls is often challenging. However, in practice it is quite likely that the prevalence sample may be a biased sample of the underlying population, or, if the case status is based on self-report, the prevalence rate thus estimated may either underestimate or overestimate the true prevalence rate. For example, in a study where case status is ascertained through a questionnaire or survey, one could argue that those who respond to the survey (i.e., responders) constitute a biased sample of the underlying population. Or, when the case status in the prevalence sample is determined by a manual chart review of patients' self-reported outcomes, one would expect bias in estimating the prevalence rate. In the Sorafenib study, the prevalence sample was a convenience sample from among the background population for which manual chart review was performed. No special selection was entailed, and we believe that it represented a sample from the same conceptual population as that of our study. No significant difference was found in tests of covariate patterns between the prevalence and background samples in further investigation reported by Heskel et al. In addition, determination of the case statuses through manual chart review is clearly more objective than the use of patient self-reports.

Finally, before claiming any findings, one must always conduct sensitivity analyses to assess the impact of biases due to violations of assumptions. If the data analysis involves testing hypotheses, it is important to discuss the direction of the bias when interpreting the findings. For example, bias toward the alternative hypothesis is more serious than toward the null because it may lead to false positive findings.

4. Discussion. The proposed case-background design has many potential applications. It may be a useful tool in epidemiological investigations where acquiring a sample of exposure data among a control group is impossible, unethical or

costly. Such situations are frequent in practice. For example, in studies of highly intimate health conditions, or criminal or stigmatized behaviors, it may be extremely hard to identify a true control group that is comparable in size to an existing case group due to ethical or privacy concerns. Further examples may be found among research topics in sociomedical sciences, population health and psychiatry.

To implement the proposed method, one needs a background sample and a prevalence sample to replace the unavailable control sample. Even though a control sample can be obtained in some of the scenarios we consider, it can be argued that the cost of the background and the prevalence samples are much lower, making the proposed design appealing. In addition, unlike the biased estimation strategy that would result from treating the background sample as an approximation to a control sample, no requirement of low disease prevalence is needed for our proposed design.

The proposed method has been demonstrated to have satisfactory robustness properties against modest violations of its key assumption: that the background and prevalence samples accurately reflect the same underlying population from which the case sample is selected. However, we emphasize that the practical guidelines provided in Section 3.4 specifying the importance of good covariate selection, consideration of biases and sensitivity analyses be followed as closely as possible when implementing this form of study.

APPENDIX: PROOF OF THE THEOREM

PROOF OF THEOREM 2.1. The estimating equation can be written as

$$\frac{1}{n_b}\sum_{j=1}^{n_b}\frac{\tilde{\boldsymbol{X}}_{bj}e^{\tilde{\boldsymbol{X}}_{bj}^T\boldsymbol{\beta}}}{1+e^{\tilde{\boldsymbol{X}}_{bj}^T\boldsymbol{\beta}}}=\frac{1}{n_p}\sum_{l=1}^{n_p}I(Y_{pl}=1)\frac{1}{n_d}\sum_{i=1}^{n_d}X_{di}. \tag{A.1}$$

We show that equation (A.1) is unbiased when $\boldsymbol{\beta}$ equals the true unknown parameter $\boldsymbol{\beta}^*$. In fact,

$$\begin{aligned}
E\left(\frac{1}{n_b}\sum_{j=1}^{n_b}\frac{\tilde{\boldsymbol{X}}_{bj}e^{\tilde{\boldsymbol{X}}_{bj}^T\boldsymbol{\beta}^*}}{1+e^{\tilde{\boldsymbol{X}}_{bj}^T\boldsymbol{\beta}^*}}\right) &= E\left(\frac{\tilde{\boldsymbol{X}}_{b1}e^{\tilde{\boldsymbol{X}}_{b1}^T\boldsymbol{\beta}^*}}{1+e^{\tilde{\boldsymbol{X}}_{b1}^T\boldsymbol{\beta}^*}}\right)\\
&= E\{\tilde{\boldsymbol{X}}_{b1}E(Y_{b1}\mid\tilde{\boldsymbol{X}}_{b1})\}\\
&= E(\tilde{\boldsymbol{X}}_{b1}Y_{b1})\\
&= \mathrm{pr}(Y_{b1}=1)E(\tilde{\boldsymbol{X}}_{b1}\mid Y_{b1}=1)\\
&= \mathrm{pr}(Y_{p1}=1)E(\tilde{\boldsymbol{X}}_{d1}\mid Y_{d1}=1)\\
&= E\left\{\frac{1}{n_p}\sum_{l=1}^{n_p}I(Y_{pl}=1)\right\}E\left(\frac{1}{n_d}\sum_{i=1}^{n_d}\tilde{\boldsymbol{X}}_{di}\right).
\end{aligned}$$

Define

$$\boldsymbol{H}(\boldsymbol{\beta}) = \int \frac{\tilde{\boldsymbol{x}} e^{\tilde{\boldsymbol{x}}^T \boldsymbol{\beta}}}{1 + e^{\tilde{\boldsymbol{x}}^T \boldsymbol{\beta}}} f(\boldsymbol{x})\, d\mu(\boldsymbol{x}), \qquad \pi_d = \operatorname{pr}(Y = 1),$$

$$\boldsymbol{\gamma} = E(\tilde{\boldsymbol{X}} \mid Y = 1), \qquad \boldsymbol{\theta} = \pi_d \boldsymbol{\gamma}.$$

We have proved in the above that $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ is a solution to $\boldsymbol{H}(\boldsymbol{\beta}) = \boldsymbol{\theta}$. Write $\boldsymbol{M}(\boldsymbol{\beta}) = \partial \boldsymbol{H}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ whose (k, l)th entry is

$$m_{kl} = \int x_{k-1} x_{l-1} e^{\tilde{\boldsymbol{x}}^T \boldsymbol{\beta}} \bigl(1 + e^{\tilde{\boldsymbol{x}}^T \boldsymbol{\beta}}\bigr)^{-2} f(\boldsymbol{x})\, d\mu(\boldsymbol{x}) \qquad (k, l = 1, \ldots, K).$$

Applying Theorem 182 in Hardy, Littlewood and Pólya (1952) to vector $\boldsymbol{x}\{e^{\tilde{\boldsymbol{x}}^T \boldsymbol{\beta}} \times f(\boldsymbol{x})\}^{1/2}(1 + e^{\tilde{\boldsymbol{x}}^T \boldsymbol{\beta}})^{-1}$, we conclude that $\boldsymbol{M}(\boldsymbol{\beta})$ is positive definite because $\operatorname{var}(\boldsymbol{X})$ is positive definite, implying that the components of $\boldsymbol{X}$ cannot be linearly dependent almost surely. The positive definiteness of $\boldsymbol{M}(\boldsymbol{\beta})$ for any $\boldsymbol{\beta}$ implies that the solution to $\boldsymbol{H}(\boldsymbol{\beta}) = \boldsymbol{\theta}$ must be unique.

Let

$$\hat{\boldsymbol{H}}(\boldsymbol{\beta}) = \frac{1}{n_b} \sum_{j=1}^{n_b} \frac{\tilde{\boldsymbol{X}}_{bj} e^{\tilde{\boldsymbol{X}}_{bj}^T \boldsymbol{\beta}}}{1 + e^{\tilde{\boldsymbol{X}}_{bj}^T \boldsymbol{\beta}}}, \qquad \hat{\pi}_d = \frac{1}{n_p} \sum_{l=1}^{n_p} I(Y_{pl} = 1),$$

$$\hat{\boldsymbol{\gamma}} = \frac{1}{n_d} \sum_{i=1}^{n_d} \tilde{\boldsymbol{X}}_{di}, \qquad \hat{\boldsymbol{\theta}} = \hat{\pi}_d \hat{\boldsymbol{\gamma}}.$$

By Theorem 8 in Hardy, Littlewood and Pólya (1952), $\hat{\boldsymbol{M}}(\boldsymbol{\beta}) = \partial \hat{\boldsymbol{H}}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ is positive definite as long as the design matrix $(\tilde{\boldsymbol{X}}_{b1}, \ldots, \tilde{\boldsymbol{X}}_{bn_b})^T$ is of rank K almost surely. By Theorem 16(a) in Ferguson (1996), $\hat{\boldsymbol{H}}(\boldsymbol{\beta}) \to_{a.s.} \boldsymbol{H}(\boldsymbol{\beta})$ uniformly on $\boldsymbol{B}_\delta = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \le \delta\}$ for some fixed $\delta > 0$, where $\|\cdot\|$ denotes the K-dimensional Euclidean distance. Since $\hat{\boldsymbol{\theta}} \to_{a.s.} \boldsymbol{\theta}$, $\hat{\boldsymbol{H}}(\boldsymbol{\beta}) \to_{a.s.} \boldsymbol{H}(\boldsymbol{\beta})$ uniformly on $\boldsymbol{B}_\delta$, we conclude that equation $\hat{\boldsymbol{H}}(\boldsymbol{\beta}) = \hat{\boldsymbol{\theta}}$ has a unique solution $\hat{\boldsymbol{\beta}}$ almost surely and $\hat{\boldsymbol{\beta}} \to_{a.s.} \boldsymbol{\beta}^*$.

By the mean value theorem,

$$\boldsymbol{0} = \hat{\boldsymbol{H}}(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\theta}} = \hat{\boldsymbol{H}}(\boldsymbol{\beta}^*) + \int_0^1 \hat{\boldsymbol{M}}\bigl(t\boldsymbol{\beta}^* + (1-t)\hat{\boldsymbol{\beta}}\bigr)\, dt \cdot (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - \hat{\boldsymbol{\theta}},$$

which yields

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = -\biggl\{\int_0^1 \hat{\boldsymbol{M}}\bigl(t\boldsymbol{\beta}^* + (1-t)\hat{\boldsymbol{\beta}}\bigr)\, dt\biggr\}^{-1} n^{1/2}\{\hat{\boldsymbol{H}}(\boldsymbol{\beta}^*) - \hat{\boldsymbol{\theta}}\}.$$

By Theorem 16(a) in Hardy, Littlewood and Pólya (1952), $\hat{\boldsymbol{M}}(\boldsymbol{\beta}) \to_{a.s.} \boldsymbol{M}(\boldsymbol{\beta})$ uniformly on $\boldsymbol{B}_\delta$, which, together with the fact that $\hat{\boldsymbol{\beta}} \to_{a.s.} \boldsymbol{\beta}^*$, yields that

$$\int_0^1 \hat{\boldsymbol{M}}\bigl(t\boldsymbol{\beta}^* + (1-t)\hat{\boldsymbol{\beta}}\bigr)\, dt \to_{a.s.} \boldsymbol{M}(\boldsymbol{\beta}^*).$$

By the central limit theorem,

$$n^{1/2}\begin{Bmatrix}\hat{\boldsymbol{H}}(\boldsymbol{\beta}^*)-\boldsymbol{H}(\boldsymbol{\beta}^*)\\ \hat{\pi}_d-\pi_d\\ \hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\end{Bmatrix}\to_d N\{\mathbf{0},\operatorname{diag}(\rho_b^{-1}\boldsymbol{\Sigma}_1,\rho_p^{-1}\Sigma_2,\rho_d^{-1}\boldsymbol{\Sigma}_3)\},$$

where

$$\boldsymbol{\Sigma}_1=\operatorname{var}\biggl(\frac{\tilde{\boldsymbol{X}}e^{\tilde{\boldsymbol{X}}^T\boldsymbol{\beta}^*}}{1+e^{\tilde{\boldsymbol{X}}^T\boldsymbol{\beta}^*}}\biggr),\qquad \Sigma_2=\pi_d(1-\pi_d),\qquad \boldsymbol{\Sigma}_3=\operatorname{var}(\tilde{\boldsymbol{X}}\mid Y=1).$$

By the δ-method,

$$n^{1/2}\{\hat{\boldsymbol{H}}(\boldsymbol{\beta}^*)-\hat{\boldsymbol{\theta}}\}\to_d N(\mathbf{0},\boldsymbol{V}),$$

where

$$\boldsymbol{V}=\rho_b^{-1}\boldsymbol{\Sigma}_1+\rho_p^{-1}\pi_d(1-\pi_d)E(\tilde{\boldsymbol{X}}\mid Y=1)\{E(\tilde{\boldsymbol{X}}\mid Y=1)\}^T+\rho_d^{-1}\pi_d^2\boldsymbol{\Sigma}_3.$$

Then, by Slutsky's theorem,

$$n^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\to_d N\bigl(\mathbf{0},\boldsymbol{M}(\boldsymbol{\beta}^*)^{-1}\boldsymbol{V}\boldsymbol{M}(\boldsymbol{\beta}^*)^{-1}\bigr).\qquad\square$$

Acknowledgments. We thank Editor Griffin, one Associate Editor and two anonymous referees for their constructive suggestions.

REFERENCES

BRESLOW, N. E. and DAY, N. E. (1999). *Statistical Methods in Cancer Research*: *Volume* 1—*The Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyons.

FERGUSON, T. S. (1996). *A Course in Large Sample Theory*. *Texts in Statistical Science Series*. Chapman & Hall, London. MR1699953

GOLDSTEIN, L. and LANGHOLZ, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *Ann. Statist.* **20** 1903–1928. MR1193318

HARDY, G. H., LITTLEWOOD, J. E. and PÓLYA, G. (1952). *Inequalities*, 2nd ed. Cambridge Univ. Press, New York. MR0046395

KEOGH, R. H. and COX, D. R. (2014). *Case-Control Studies*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **4**. Cambridge Univ. Press, Cambridge. MR3443808

KUPPER, L. L., MCMICHAEL, A. J. and SPIRTAS, R. (1975). A hybrid epidemiology study design useful in estimating relative risk. *J. Amer. Statist. Assoc.* **99** 832–844.

LIDDELL, F. D. K., MCDONALD, J. C., THOMAS, D. C. and CUNLIFFE, S. V. (1977). Methods of cohort analysis: Appraisal by application to asbestos mining. *J. Roy. Statist. Soc. Ser. A* **140** 469–491.

MIETTINEN, O. S. (1976). Estimability and estimation in case-referent studies. *Am. J. Epidemiol.* **103** 226–235.

NURMINEN, M. (1989). Analysis of epidemiologic case-base studies for binary data. *Stat. Med.* **8** 1241–1254.

PRENTICE, R. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73** 1–11.

PRENTICE, R. and BRESLOW, N. E. (1978). Retrospective studies and failure time models. *Biometrika* **65** 153–158.

PRENTICE, R. L. and PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66** 403–411. MR0556730

ROBERTS, L. R. (2008). Sorafenib in liver cancer—Just the begining. *N. Engl. J. Med.* **359** 420–422.

SELF, S. G. and PRENTICE, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.* **16** 64–81. MR0924857

DEPARTMENT OF POPULATION HEALTH SCIENCE
AND POLICY
ICAHN SCHOOL OF MEDICINE AT MOUNT SINAI
1425 MADISON AVENUE
NEW YORK, NEW YORK 10029
USA
E-MAIL: john.spivack@mountsinai.org

DEPARTMENT OF BIOSTATISTICS
MAILMAN SCHOOL OF PUBLIC HEALTH
COLUMBIA UNIVERSITY
722 WEST 168TH STREET
NEW YORK, NEW YORK 10032
USA
E-MAIL: bc2159@cumc.columbia.edu